

PDF hosted at the Radboud Repository of the Radboud University Nijmegen

The following full text is a publisher's version.

For additional information about this publication click this link.

<http://hdl.handle.net/2066/149175>

Please be advised that this information was generated on 2017-12-05 and may be subject to change.

OPINION

Open Access



Epigenome data release: a participant-centered approach to privacy protection

Stephanie O. M. Dyke^{1*}, Warren A. Cheung^{2†}, Yann Joly¹, Ole Ammerpohl³, Pavlo Lutsik⁴, Mark A. Rothstein⁵, Maxime Caron², Stephan Busche², Guillaume Bourque², Lars Rönnblom⁶, Paul Flicek⁷, Stephan Beck⁸, Martin Hirst⁹, Henk Stunnenberg¹⁰, Reiner Siebert³, Jörn Walter⁴ and Tomi Pastinen^{2*}

Abstract

Large-scale epigenome mapping by the NIH Roadmap Epigenomics Project, the ENCODE Consortium and the International Human Epigenome Consortium (IHEC) produces genome-wide DNA methylation data at one base-pair resolution. We examine how such data can be made open-access while balancing appropriate interpretation and genomic privacy. We propose guidelines for data release that both reduce ambiguity in the interpretation of open-access data and limit immediate access to genetic variation data that are made available through controlled access.

Sequencing-based techniques such as integrative transcriptomic measurements of gene expression and epigenomic measurements of chromatin structure are increasingly applied to the study of genome function. Open sharing of human epigenome data is of great importance to progress in the large-scale data-intensive biomedical research carried out by the International Human Epigenome Consortium (IHEC), of which we are members. Data-sharing facilitates subsequent research, enhancing reproducibility and the translation of research into new knowledge of health and disease.

Evidence suggests that genetically mediated variation within human tissues is abundant, easily mapped and shared between tissues [1]. From a genomic privacy standpoint, DNA sequence information can lead to the re-identification of research participants' data by genetic

matching — this has been referred to as 'attribute disclosure attacks using DNA' (ADAD) [2]. Here, we discuss the current practices and privacy protections currently available for the release of genomic and related data. We quantify the extent to which identifying DNA sequence information confounds anonymization using the example of methylation data, and conduct an ethical-legal analysis of the issues raised with respect to the privacy and autonomy of research participants. Finally, we propose open-access data-release policies to address these issues.

De-identification of data by removing direct identifiers (such as participants' name, date of birth, social insurance numbers and facial images) is widely used for shared research data. In North America, anonymization implies that the de-identified data are no longer linked to any identifiers. By contrast, coding refers to an alphanumeric 'code' that links de-identified data to identifiers. In this analysis, we draw a distinction between the re-identification of data — its attribution to an individual by matching identified (named) genetic information to anonymized data — and the potential to link two anonymized datasets. Absolute anonymization of even small amounts of DNA sequence information can be impossible given the extent to which DNA sequence is unique to individuals [3, 4], but epigenomic data lend themselves more readily to anonymization.

When there is a reasonable risk that data can be re-identified, or there are limitations on the use of the data in different types of analyses, another strategy to enable the data to be shared is to control access to it. 'Controlled access' ('managed access') has generally been applied to data types that provide extensive DNA sequence information from an individual. Researchers must apply for access to such datasets and be approved by a 'Data Access Committee' (DAC). The ability to re-identify and misuse research data is considered less likely when the data are shared under controlled access arrangements that involve a review of applicants'

* Correspondence: Stephanie.Dyke@mcgill.ca; Tomi.Pastinen@mcgill.ca

†Equal contributors

¹Centre of Genomics and Policy, Department of Human Genetics, McGill University, Montreal, QC H3A 0G1, Canada

²Department of Human Genetics, McGill University and Genome Quebec Innovation Centre, Montreal, QC H3A 0G1, Canada

Full list of author information is available at the end of the article

credentials, a review of their research plans, verification that the proposed research has been approved by an ethics committee or that a waiver has been obtained, and the signing of a contract referred to as a Data Access Agreement that forbids (amongst other things) the re-identification of data. DACs can also provide some degree of post-authorization oversight of data use [5]. These measures can, to varying degrees, limit data access and analysis, so they have been perceived by some members of the research community as hindering ‘crowd-sourcing’ or collaborative analysis of publically funded genomic datasets [6]. Other concerns include delays that result from the controlled access process and its lack of transparency [7].

Numerous security strategies can increase the level of protection of data (for example, firewalls or encryption) or enhance privacy (for example, iDASH [8] and Bio-PIN [9]). Typically though, data security measures serve to reinforce controlled access distribution and do not address its main limitations: restricting acceptable data use and aggregation. An emerging approach to providing broad access to data while protecting the privacy interests of research participants is that of data ‘safe havens’ — protected IT environments for pooling data (such as DataSHIELD [10]). The strengths of this approach are that it aims to reduce the risks of distributing large amounts of data to individual researchers and decreases reliance on contracts and other legal protections that are neither fail-proof nor evenly provided internationally, and which can be difficult to enforce.

Following the model of the National Institute of Health (NIH) Roadmap Epigenomics project, an IHEC partner, processed IHEC epigenomes are publically accessible in appropriate data archives, track hubs or similar summary data formats. Associated raw sequence data and metadata information are also shared, either through open-access or controlled-access mechanisms. Similarly, The Cancer Genome Atlas (TCGA) provides publically accessible ‘Level 3’ summarized methylation calls, whereas controlled access to ‘Level 1’ and ‘Level 2’ data restricts the availability of raw sequence and mutation calls [11]. Open-access data, which are freely available for anyone to use, typically include intensities of signal (such as gene expression or DNA–protein interaction) or levels of methylated cytosine. Such summary data do not report genetic variation directly, and their release reflect the strategies developed for the open-access release of array-based gene expression data by the National Centre for Biotechnology Information (NCBI) Gene Expression Omnibus (GEO) or the European Bioinformatics Institute (EMBL-EBI) ArrayExpress (AE) databases. Users must rely on the data submitter for appropriate processing of data, potentially leading to biological misinterpretation.

DNA methylation data are an example of a form of epigenetic information that can lead to misinterpreted results because of the presence of genetic variants, given its reliance on CpG (cytosine-phosphate-guanine) dinucleotide contexts (CpGs) as the unit of information. Other components of epigenome mapping data (such as DNase hypersensitivity sites or chromatin marks [12–14]) also show evidence of genetic governance, but the density of these traits and how they are shared across tissues has only been studied in smaller datasets. Bisulfite conversion causes unmethylated cytosines to be converted to uracil, allowing methylated and unmethylated cytosines to be distinguished. Whole-genome bisulfite sequencing (WGBS) is a high-throughput, genome-wide DNA methylation interrogation technique that reports methylated and unmethylated cytosines at CpG sites within a reference genome.

WGBS is biased at the start and end of reads because it includes unmethylated cytosines that are added during overhang repair and 5′ underconversion from adapter re-annealing [15, 16]. It also confounds methylated cytosines and hydroxymethylcytosines, which are of particular importance in certain cell types (for example, in the nervous system) [17, 18]. We focus on genetic confounders: WGBS additively measures the frequency of cytosines in CpH (cytosine-phosphate-(non-guanine nucleotide)) contexts, as well as thymine polymorphisms.

Case study: genetic information in methylation data

Strand-specific WGBS measures CpG methylation for the forward and reverse strands independently, but both strands usually have concordant methylation rates. Nevertheless, when the cytosine of the CpG is mutated to adenine or guanine on the forward strand, asymmetric methylation rates are measured (Fig. 1). When the cytosine is mutated to thymine, all reads are counted, but forward reads that contain the thymine mutation are miscounted as bisulfite-converted unmethylated cytosines, and reverse reads measure CpH methylation at the mutated site. In both cases, the polymorphism can be detected by the base-paired genetic variation in reverse reads [19, 20] or externally by direct genome sequencing or genotyping arrays.

We identified genomic CpGs from WGBS in which the measured methylation rate is due to genetic rather than epigenetic variation and is independent of tissue type (Fig. 2). We did this by filtering for CpGs that have a static methylation rate in all tissues from the same individual in the NIH RoadMap Epigenomics (Roadmap) [21] WGBS samples (Additional file 1: Table S1) but which vary between individuals. A total of 5.9 million candidate CpGs were identified from a pool of 24 million well-measured CpGs present in most of the

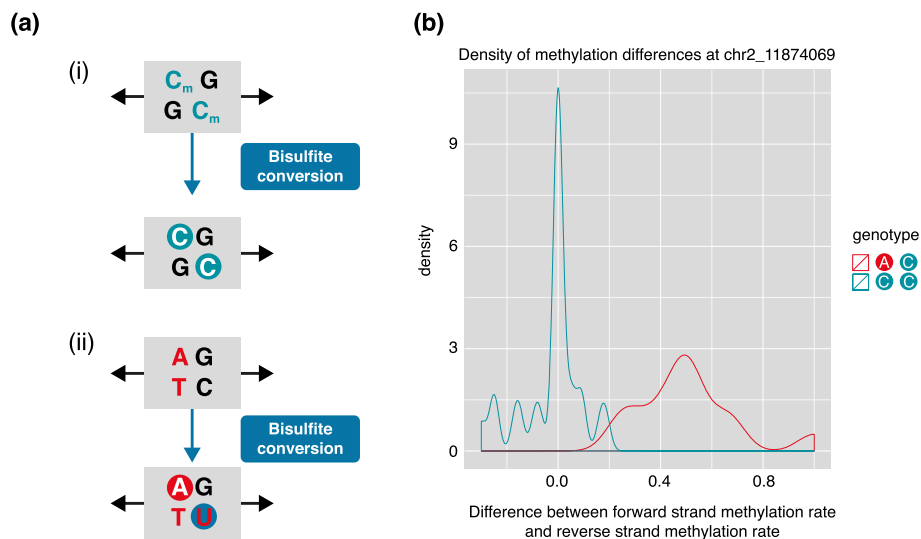


Fig. 1 Genotypic differences in forward and reverse strand methylation. **a** (i) On reads from both strands of the wild-type C allele, the methylated C usually remains as C after bisulfite conversion, and is counted as methylated. This results in a mean difference of methylation between the strands of 0. (ii) For the allele where the methylated C is replaced by A, reads on the forward strand have the A at the CpG site and are not counted, whereas the reads on the reverse strand have the C bisulfite-converted to U and are counted as unmethylated. This results in a mixture of methylated and unmethylated reads on the reverse strand, whereas there are only methylated reads on the forward strands. **b** Heterozygotes that have A and C alleles (red) are compared with homozygotes that have two copies of the C allele (turquoise). We see negligible difference in methylation rate between forward and reverse strands in the 26 homozygous individuals, but an average of around 50 % more methylation on the forward versus the reverse strand in the 13 heterozygous individuals

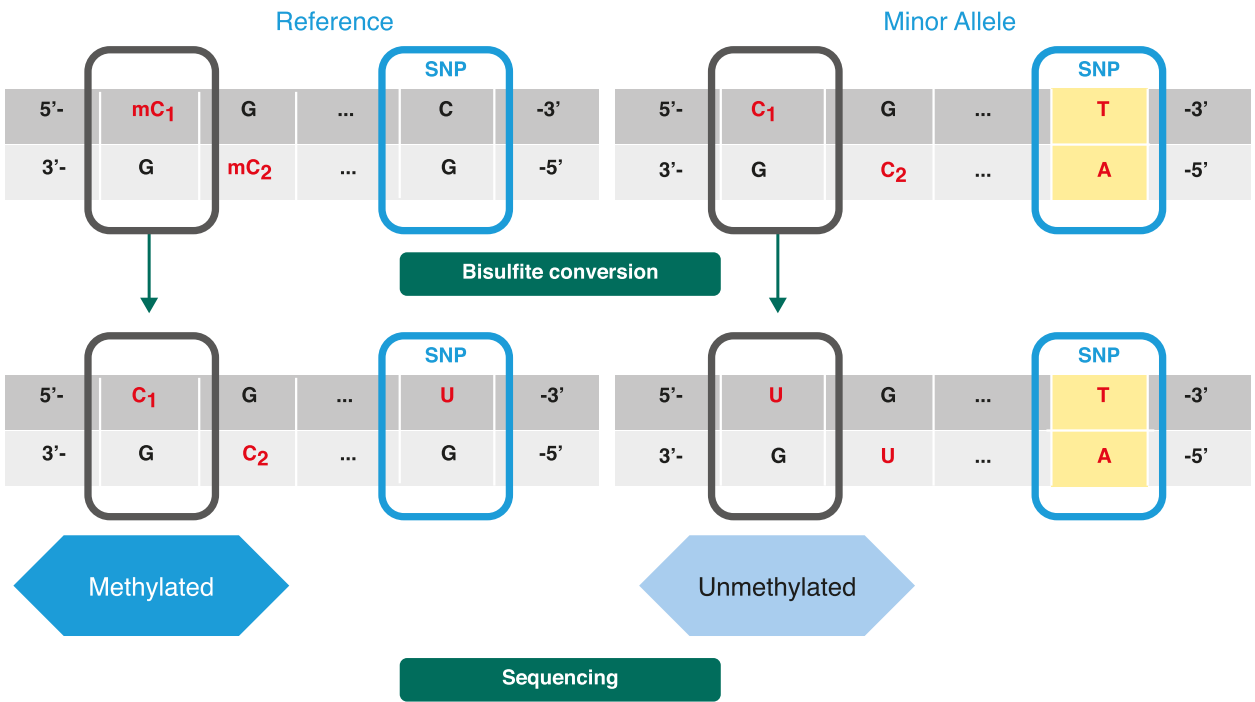
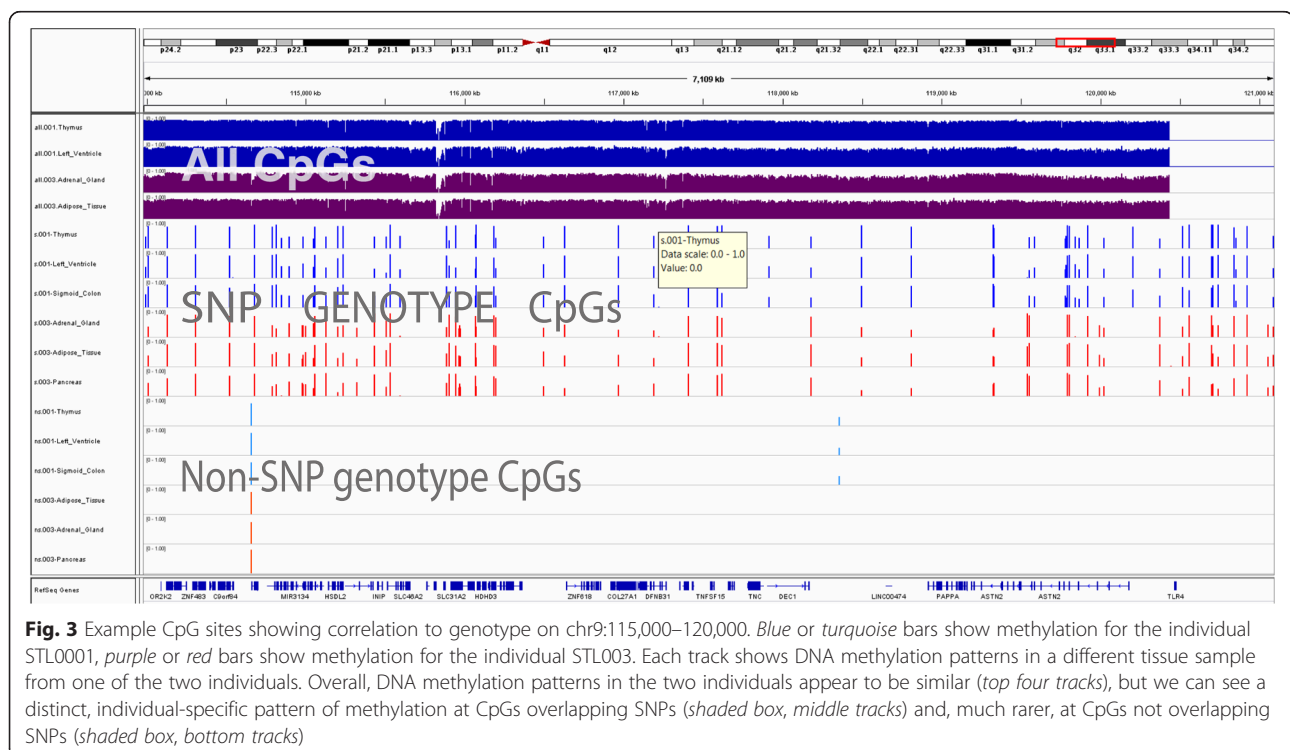
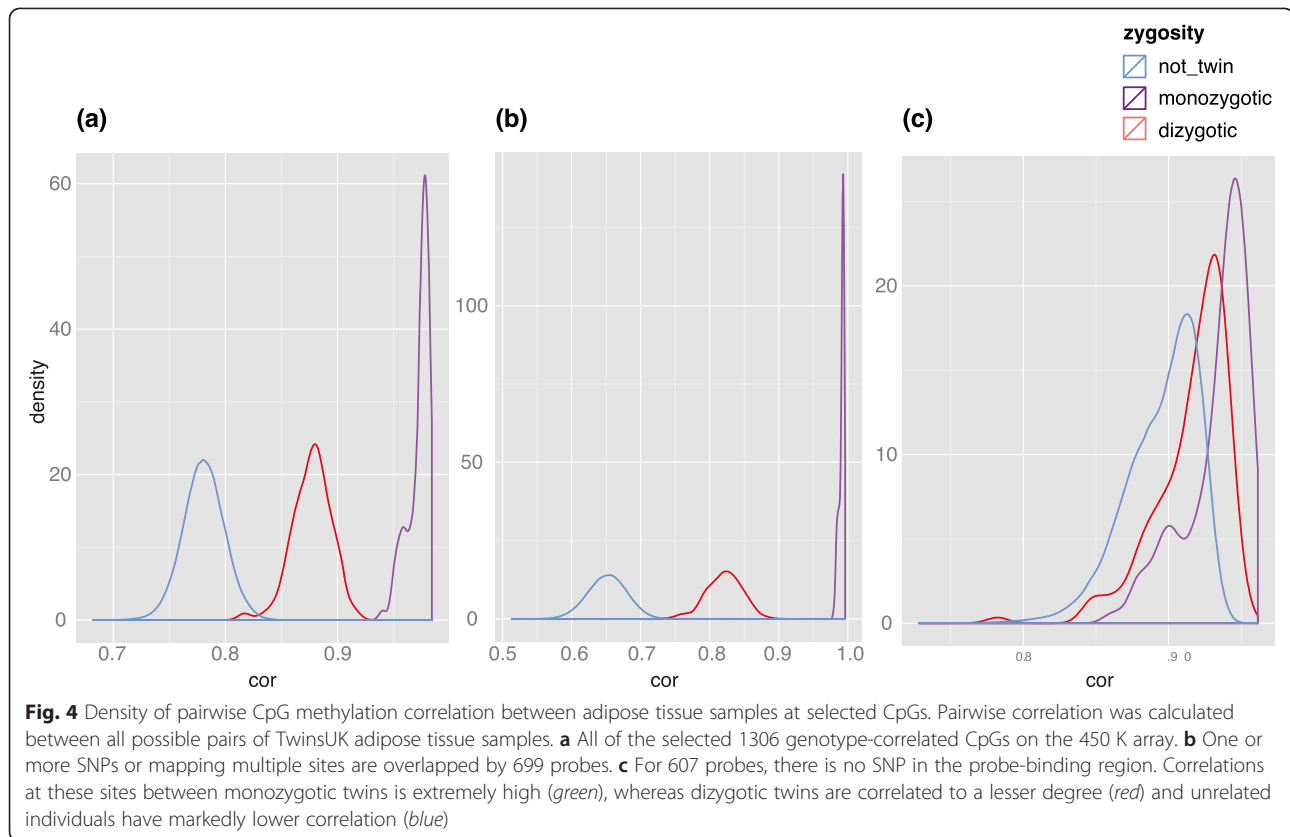


Fig. 2 Example in which methylation is indirectly affected by a SNP. The CpG site is normally methylated (left) when the genomic sequence at a downstream SNP position is a C. When the downstream SNP is mutated to a T, the CpG site is affected and becomes unmethylated, allowing the conversion of the cytosine residues at the CpG site to uracil (right)

Public WGBS datasets therefore contain thousands of genetic variants, predominantly known common variants, that disrupt CpGs. Other sites that show high variability among individuals, but not tissues, may be subject to indirect genetic effects or may contain rare variants. We validated the Roadmap/EMC-identified highly predictive genotype-correlated CpGs using independent methylation and genotype sequencing data from adipose tissue [23]; only CpGs overlapping a

Other methylation interrogating techniques also expose genetic information. The Illumina Infinium Human-Methylation 450 K BeadChip Array (450 K) provides genome-wide microarray interrogation of 485,577 CpG targets. We identified probes from public domain 450 K data that had a static methylation rate in all tissues from the same individual but which had variable methylation rates between individuals. After excluding all 65 SNP-targeting 'rs' probes, 1306 'cg' probes (Additional file 2: Table S2) matched leukemia cancer and normal cells by genotype [24]. When validated in adipose tissue [25], these probes showed extremely high correlation in monozygotic twins compared with that in dizygotic twins and unrelated individuals (Fig. 4).





Removing direct genetic variation

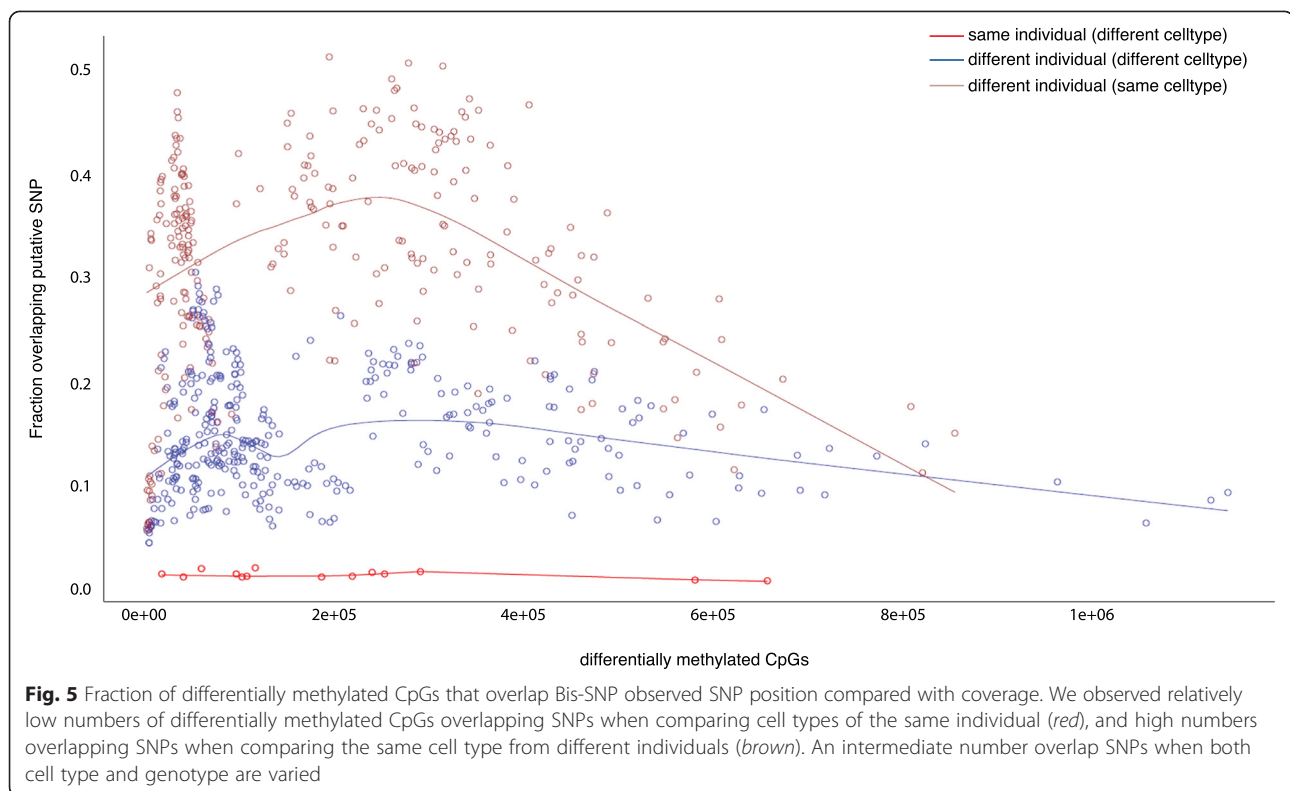
The strand-specific WGBS approach allows unequivocal distinction between genetic and epigenetic variation through direct sequencing of base-paired nucleotides at the same position as the variation on the opposing strand. Using Bis-SNP [19], we identified the genotype of reference CpG sites from normal purified blood WGBS datasets de novo (without dbSNP information), validating against heterozygotes detected in genotyping arrays. We identified 66.5 % of arrayed variants at CpGs, reducing the fraction of CpG sites that contained variants from 11 % to 3.7 % of the genotyped CpGs. Of the genotyped positions overlapping CpGs, 0.029 % were incorrectly called (false positive SNP or incorrect variant called). Low coverage (<10 reads) contributed to the vast majority of the mislabeling.

Using SNP frequencies from dbSNP137, a median of 95 % of covered reference CpG positions in the WGBS data were retained after removing detected SNPs and unclear cytosine contexts. Detection of variants at genotyped CpGs was increased to 75 %, and erroneous SNP calls were reduced to 0.024 %. When focusing on high-coverage CpG sites (with a minimum of 15× coverage), we identified 0.4–1.5 % (median 1.3 %) of high-coverage CpGs per sample as having SNPs (samples had 1 million to 20 million high-coverage CpGs, median 5.7 million).

We next examined differentially methylated CpGs (methylation rate difference >30 %) between pairs of samples. Overall, between 1 % and 50 % of the differentially methylated cytosines (median 20 %) were identified as overlapping sequence variants in one or both samples (Fig. 5). When comparing the same blood cell type between different individuals, an extremely large fraction (up to 50 %) of differentially methylated CpGs were due to SNPs (median = 33 %). By contrast, samples from different cell types of the same individual (16 pairs in total from 7 individuals) showed a median of 1.5 % overlap with SNP calls, indicating that differential methylation at heterozygous sites is rare. Varying both tissue and genotype, SNPs had an overall intermediate contribution to the differential methylation (median 14 %) at CpG sites, indicating that while CpGs that have true differential methylation were detected (above the intra-tissue rate), genetic variation at the CpG site remained a substantial influencing factor.

Vulnerability of metadata

There remains a very small risk of re-identification of research participants by matching their identified named genomic information to data from a study participant. We therefore consider the consequences of potential re-identification of associated clinical/healthcare information



and other lifestyle or demographic information, which may be studied and available from metadata and study parameters. Some of these metadata may also increase the likelihood of re-identification of the dataset.

Epigenome mapping projects include samples from a number of population cohorts with varying health conditions, including rare diseases. It is clear that the epigenome is impacted by disease state; therefore, some categorization of the health status of the donor may be necessary depending on the tissue studied. The use of controlled vocabulary with disease ontologies (such as the NCI Metathesaurus used by IHEC) allows for this information to be reported in a standardized manner, which reduces the risk of inadvertent disclosure of more detailed health information if a dataset were to be re-identified. Nevertheless, some medical information does not correspond neatly to existing ontology terms and it may be necessary to allow for additional ‘free-form’ text relating to disease and donor health status.

For individuals with a rare disease or other rare phenotype, disease or donor health status information could potentially increase the risk of re-identification of epigenomic data in the same way as seemingly innocuous ‘demographic’ information. For example, full date of birth and place of residence have been shown to enable re-identification of healthcare data in some circumstances

[26]. Information on rare disease status can increase the risk of re-identification not only because rare diseases are rare, but also because the disease often presents outwardly visible characteristics that could link a whole dataset more rapidly to an individual. Furthermore, some rare diseases imply potential carrier status for relatives and the disease may also be associated with potentially stigmatizing information. For example, bilateral striopallidodentate calcinosis, with fewer than 200 known cases and for which the genetic basis is not fully understood (familial and sporadic forms, genes unknown) may cause personality changes and dementia [27]. Mental health information is generally considered to be stigmatizing and it is often provided special protection by law [28]. Severe conditions such as this are, however, unlikely to be kept private once symptomatic, so the main risk is the increased likelihood of re-identification of other information in the dataset.

Rare disease information may also reveal an individual’s likely ancestry or geographical location. For example, Tay-Sachs disease has a higher prevalence in individuals of Ashkenazi Jewish descent [29], and Leigh syndrome in the Saguenay-Lac-Saint-Jean region of Quebec [30]. In some cases, such associations may result in a loss of privacy. Furthermore, the experience of projects in which rare disease genetics data have been shared indicates that patients and their families are

willing to accept voluntarily the risks associated with potential re-identification if they have been explained to them. While this acceptance of risk may not be greater than in other research circumstances, it can be presumed that there are greater expectations of benefits from involvement in rare disease research. We propose points to consider for assessing the risk of sharing rare disease information in open-access data sets (Table 1). These relate to the potential for re-identification, the privacy and sensitivity of rare disease data, and research participants' consent.

While it is very difficult to quantify the likelihood of re-identification in these cases, a 'rarity' threshold for point 3, for example, could be considered that would be relative to the availability of information on place of residence and the visibility of the disease (points 1 and 2). If the answer to point 4 or 5 is yes, we recommend holding rare disease information in 'controlled access' while clearly indicating its availability.

Most current epigenome mapping projects focus on the characteristics of human cell types or tissues and de-identification is the norm. Nevertheless, datasets commonly include two other important categories of metadata — donor age and ethnicity — which impact interpretation of the data and are therefore important to share as openly as possible [31, 32]. The risk of re-identification of anonymized datasets from 'demographic' metadata requires project-specific consideration, depending mainly on other sources of available information and on the group sizes of a given demographic [26]. Standards, such as the US Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule, significantly decrease re-identification risk (for example, by using age, not date of birth, with a category for ages over 90 years) [33].

For ethnicity, the risk mainly applies to minority groups, with the re-identification risk varying (similar to that for rare disease metadata). Ethnic origin or ethnicity is included as a surrogate marker for genetic similarity or relatedness in order to improve the quality of research

results in terms of their significance generally and for individuals [34–36]. This metadata use creates difficulties with respect to adopting publically acceptable group designations [37]. Given the diversity of approaches for recording ethnicity (or not) in different parts of the world, and the benefits of standardizing descriptors in research, consulting local census categories and assigning a limited set of choices based on the populations studied would help in addressing social and political issues that might affect research participants [38]. However, populations requiring special attention, such as small ethnic groups that may be more prone to the risks of re-identification, need to be identified as such if their data are to be shared with extra protections. This can lead to a quandary as census categories may purposely avoid asking for this information. We suggest reviewing lists of proposed descriptors for sample populations, and, if possible, providing preset lists to select from that are based on locally acceptable designations such as those of national census categories. For small or vulnerable populations, the determination of which will also usually depend on local context, we also suggest moving this information (and potentially other data from these individuals) to the 'controlled access' portion of the data.

Mitigating risk for data release

Anonymized genome-wide DNA sequence information that is contained within public repositories can be linked to individual participants [2]. The main reason this has not prevented its public release in some circumstances (for example, with appropriate consent and following an assessment of the sensitivity and identifiability of associated metadata) is that, in the vast majority of cases, to do so would require access to an individual's identified genetic data from another source, in which case the information, health-related or otherwise, that it contains would probably not be protected. Anonymized genome-wide genetic data can also sometimes be re-identified by other routes, such as through surname inference for well-documented collections [39]. Furthermore, for functional genomic data (such as RNA-expression profiles), considerable efforts would be required to match datasets by tissue of origin and processing techniques. This has been studied for gene expression arrays using pre-existing knowledge of genetic variation that impacts gene expression differences in populations [40] and is a much more complex route to a privacy breach [2].

Open-access DNA methylome data contains DNA-sequence information that could potentially be used as re-identifying information through genetic matching. However, the majority of genotype-resolving CpGs in WGBS data directly overlap known SNPs, representing other sequence contexts misleadingly released in CpG-methylation tracks. The CpGs disrupted directly by SNPs that are currently present in open-access epigenome

Table 1 Points to consider when sharing rare disease information

Points to consider
1 Is the place of residence provided (even indirectly, for example, in the project name)?
2 Is the rare disease outwardly visible?
3 How rare is the disease?
4 Does the rare disease provide information about the likely geographical location of individuals?
5 Does the rare disease provide information about ethnicity that may be considered potentially stigmatizing?
6 Was the participant aware of the potential risks of data re-identification?

data resources can be efficiently removed from high-coverage data by pre-filtering prior to release using existing algorithms or genotyping resources, with minimal loss of 'true epigenetic' information. Over 75 % of the disrupted CpGs could be eliminated with nearly 0 % erroneous calls, affecting only 1.5 % of the methylome. The genotypically resolved raw datasets would still allow interrogation of these disrupted CpGs, and in cases such as cancer genomes, somatic mutations could be reported while keeping germline mutations under controlled access (as in the TCGA policy [11]). Unfortunately, filtering cannot be used as effectively for all data types, including that generated by non-stranded bisulfite-sequencing methods (such as post-bisulfite adaptor tagging (PBAT) [41]) and methylation array data. Nevertheless, the effects of common genetic variation could still be reduced by masking sites (CpGs or probes) that have common SNPs [42, 43]. Methylation data with direct genotype variation removed would have, in our view, very low re-identification risk, probably in the same order as that for functional genomic data. For summary-level open-access data (where the user cannot reprocess the reads), such steps should precede deposition to public archives or availability in public track hubs by data producers. Patterns of data omission resulting from variants at CpGs, the presence of undetected genetic variation, and the proven existence of strong indirect (non-CpG disrupting) genetic effects on methylation within the same tissue [1] all indicate that residual genetic information will remain within methylome profiles. We have therefore also proposed additional measures to mitigate the impact of this very remote potential re-identification risk because we see great value in openly sharing the associated health and disease information and information on age and ethnicity.

Generally speaking, the greater the likelihood of re-identification and the greater the possibility that harm may occur as a result of re-identification, the greater the precautions and safeguards ought to be. For health-related and other private information, it would not be safe to assume that individuals would not generally feel distressed and would not suffer from stigma, if not discrimination, if this information were to become widely available. The 'reasonableness standard' determines that only information that can reasonably be expected to identify an individual is generally considered personal or protected by privacy laws and is included in many laws and conventions addressing data protection [44]. Following this standard, our position is based on careful evaluation of the reasonable likelihood that the data might lead to re-identification of participants. A similar approach has been taken in other large-scale data sharing collaborations such as the International Cancer Genome Consortium [45]. Furthermore, the level of privacy we feel we should strive

for is one at which both the likelihood of re-identification and any potential resulting harm are very low. This level of risk is justified in light of the public benefits of research, better understanding of health and disease, and better preventative, diagnostic, prognostic and treatment strategies that may result from epigenetic research. Our strategy relies on responsible data preparation and can benefit from additional 'Points to Consider', such as those proposed in Table 1, for assessing rare disease information.

Although documented incidents of discrimination or stigmatization on the basis of genetic information are largely limited to highly hereditary Mendelian disorders, these rare incidents have generated substantial media coverage and significant public concern [46, 47]. Several studies demonstrate that anxiety over genetic discrimination deters people from participating in promising research projects and even from undertaking clinically relevant genetic testing, even when anti-discrimination legislation has been in place for many years [48–51]. Misperception could be attenuated by providing more accessible information on privacy and anti-discrimination protections and their limitations, and a more balanced account of occurrences of genetic discrimination. Individuals might also be willing to accept the low risk of re-identification if the risks and benefits of the research are carefully explained and researchers pledge to protect the confidentiality of information to the extent possible. Information about data sharing and its risks ought to be provided during the consent process, as even consent to the broad research use of data may not be understood by participants as also implying consent to the widespread international sharing of data. This presents challenges as the risks or method of data sharing may not be known in advance. Representations of absolute protection should be avoided. Participants should also be informed that the sharing of health and other information via social media and other internet platforms may allow them to be matched to their anonymized research data. Such a patient/participant-centered approach would be respectful of participant autonomy and dignity, focusing on education and transparency, and not promising unrealistic levels of protection. The Personal Genome Project (PGP) pioneered a route for openly sharing integrated genomic, environmental and medical or trait data [52] in 2005, which was subsequently implemented in four countries (USA, Canada, UK and Austria). PGP successfully addressed many issues using an innovative open consent protocol [53]. Despite the explicit risk of re-identification, only 3.8 % of participants have withdrawn from the PGP over the past 10 years [54], suggesting high levels of participant acceptance and low levels of adverse risk from openly shared data.

Numerous regional and national laws have been enacted to protect individuals from undesired use of

their medical and genetic information, particularly from genetic discrimination in insurance and employment [55]. Nevertheless, it is currently unclear whether genetic discrimination legislation would apply to all kinds of epigenetic data because of the definitions of genetic data used in such legislation [56, 57]. For example, the US Genetic Information Nondiscrimination Act, 2008 (GINA) probably would not apply to epigenetic information since under this law the definition of a genetic test is limited to ‘an analysis of human DNA, RNA, chromosomes, proteins, or metabolites, that detects genotypes, mutations or chromosomal changes’ [58]. The German law ‘Gendiagnostikgesetz’ presents a similar situation as it defines in its §3 a genetic test as a directed test to diagnose the ‘genetic characteristics’ of a person. ‘Genetic characteristics’ are defined as ‘inherited or in between conception and birth acquired, human-derived genetic information’. In the US, the enactment of the Affordable Care Act of 2010 provides important protections against genetic discrimination in health insurance because it prohibits the denial of coverage or other adverse treatment on the basis of any preexisting health conditions or health information. Thus, this law goes beyond GINA (which only applies to asymptomatic individuals) in ensuring nondiscrimination against affected individuals in health insurance coverage. In addition, requirements for ethics review of research provide additional protection in many jurisdictions.

More robust privacy and anti-discrimination laws may be needed at the national level to efficiently address epigenetic discrimination without unduly restricting the flow of research data. However, these concerns reach beyond the context of ‘OMICS’ research. Society may have to re-conceptualize and contextualize medical confidentiality and personal privacy so that they remain relevant in the context of information technology developments and the sharing of health information through social media and the World Wide Web [59]. As demonstrated by PGP [54] and advocated by the Global Alliance for Genomics and Health, we believe it is possible to reconcile privacy protection and the protection of public benefits from scientific research that uses personal information by carefully examining the risks and using tailored data-release strategies.

Epigenomic data may also convey health-related and environmental information directly (for example, history of cigarette smoking). Discussion of these issues has been initiated [56, 60], but beyond the known impacts of smoking, alcohol consumption, chronological age and certain diseases (predominantly cancers), which are often known at sampling, epigenetic signatures for environmental exposures or disease risks have not matured sufficiently to allow assessment of their impact on data-sharing practices.

Removal of direct genotype information in methylome analyses mitigates substantial re-identification risks. Confident re-identification on the basis of the remaining methylome and other open-access epigenomics data would probably require considerable efforts. While absolute privacy cannot be guaranteed with high-throughput genomic data, we have outlined a consistent approach that limits the risks associated with open-access metadata release, aiming to allow categorization of data (for example, epigenome from normal or diseased tissue) rather than performing in-depth phenotypic correlations. Ideally, solutions that provide the benefits of open-access sharing while protecting the interests of research participants will be developed. Simultaneously, efforts to improve controlled-access mechanisms and processes for granting informed consent should be pursued. These include developing standard consent information materials and data-access agreements, and streamlining and further simplifying processes for the approval of data access.

Methods

CpG site analysis from Roadmap Epigenomics WGBS data

We tested CpG sites reported in the NIH Roadmap Epigenomics datasets in the following manner. To assess sites for intra-individual variation, we considered only sites with measurements in at least three samples from the same individual, and we computed the standard deviation of the methylation at the interrogated site. We required over half of the individuals (three out of the five) to have a standard deviation less than 0.07 at this site (bottom 70 % in a test of 100,000 CpG sites). We filtered for a minimum level of inter-individual variability by requiring the range of the methylation among the samples to be at least 15 % (top 35 % in a test of 100,000 CpG sites).

Internal assessment of genotype-methylation correlation

Genotypes for the samples were obtained using Illumina 2.5 M and 5 M genotyping arrays. For each CpG site, we correlated the methylation at this site against all SNPs within 10 kb. We modeled a linear relationship between the genotype at the SNP site and the methylation rate at the CpG site. This views each allele for the SNP as having an associated methylation rate for the CpG site, and the overall methylation rate at the CpG site as being the average of the methylation rates of the SNP alleles present in the individual. For each CpG-genotype pair, we use the fitted slope and intercept across all available samples to extrapolate the best-fit mean methylation rate for each of the three genotypes. To predict the genotype for a given methylation level, we selected the genotype with methylation rate closest to the observed methylation level.

Determination of genotype from WGBS and detection of mislabeled epigenetic variation

Bis-SNP 0.82.2 [19] was applied to the aligned and filtered reads of the purified blood samples to call SNPs directly from the strand-specific sequencing data. We limited our analysis to samples with at least 10× average read coverage (24 samples with read coverage from 12× to 22×, interrogating an average of 254,000 sites per sample). We first applied Bis-SNP without providing any prior variation information from dbSNP, evaluating all sites under the worst-case assumption of rare SNPs with no prior information. Genotype of CpG-context-altering heterozygous SNPs were determined using the Illumina 2.5 M genotyping array. Genotypes extracted using Bis-SNP without prior dbSNP frequency were compared against genotyped reference CpG sites to determine the ability to detect true heterozygous mutations as well as the rate of CpGs that were falsely identified as mutated.

We subsequently investigated the prevalence of sequence variation in methylation data by running Bis-SNP using the SNP frequency information from dbSNP137, and by examining sites with substantial read coverage ($\geq 15\times$) and large differences in methylation between samples ($>30\%$).

Roadmap Epigenomics WGBS data

Processed graphs of methylation proportions aligned to hg19 from Roadmap Epigenomics WGBS datasets were downloaded from the NCBI GEO repository [61]. We considered samples when multiple tissues were available from the same individual, a total of 49 tissue samples across five individuals (Additional file 1: Table S1). Samples were processed for bisulfite-converted methylation sequencing as described by Lister et al. [62]. CpG sites that had at least four reads (combining reads on both strands) were reported.

McGill epigenome mapping centre datasets

We assessed the correlation between methylation and genotypes in seven projects spanning tissues from naïve T cells (11 samples), cortical and trabecular bone (3 samples), muscle (7 samples), purified blood (29 T-cell, 20 monocyte and 7 B-cell samples) and whole peripheral blood (6 samples), crushed bone (3 samples), and adipose tissue (8 samples) (97 samples in total). Sequencing data are available through the McGill Epigenomics Mapping Portal [22]. Raw data are available through EGA under the study “McGill Epigenomics Mapping Centre” [EGA: EGAS00001000995].

We used the subset of the purified blood samples obtained from 28 normal Swedish individuals to evaluate genetic variation that had been mislabeled as epigenetic

differences. A total of 37 samples were analyzed from the three purified blood cell populations (CD14⁻ CD4⁺ T-cell samples, CD14⁺ monocyte samples and CD19⁺ B-cell samples).

DNA extraction

Genomic DNA (gDNA) was isolated using the NORGEN purification kit (Norgen Biotek Corporation, Canada) according to the manufacturer's protocol. All quantifications were carried out using Quant-iT PicoGreen (Life Technologies, Burlington, ON, Canada).

Whole-genome shotgun bisulfite sequencing

WGBS gDNA library preparations were carried out using the TruSeq DNA Sample Prep Kit v2 (Illumina) with an added bisulfite conversion step. gDNA (1–3 µg) spiked with 0.1 % (w/w) unmethylated λ DNA (Promega, Madison, WI, USA) was fragmented to 300–400 bp peak size using the focused-ultrasonicator E210 (Covaris, Woburn, MA, USA) to generate double-stranded DNA with 3' or 5' overhangs. Fragment size distribution was controlled on a Bioanalyzer DNA 1000 Chip (Agilent, Mississauga, ON, Canada). End repair, sample purification with AMPure beads (Beckman Coulter, Mississauga, ON, Canada), adenylation of 3' ends, and adaptor ligation was carried out as per Illumina's recommendations. The ligation product was cleaned up by one AMPure purification step, the purified DNA then analyzed on a Bioanalyzer High Sensitivity DNA Chip (Agilent), and quantified by PicoGreen before undergoing bisulfite conversion using the Epitect Fast DNA Bisulfite Kit (Qiagen, Toronto, ON, Canada) according to the manufacturer's protocol. Bisulfite-converted DNA was quantified using OliGreen (Life Technologies), and based on quantity amplified by four to six cycles of PCR using the Hifi Uracil + DNA polymerase (Kapa Biosystems, Woburn, MA, USA) according to the manufacturer's protocol. Amplified libraries were validated and quantified on Bioanalyzer High Sensitivity DNA Chips and underwent 100 bp paired-end sequencing on Illumina HiSeq2000 or HiSeq2500 systems.

Generated reads were aligned to the bisulfite-converted reference genome using the Burrows-Wheeler Alignment tool (BWA). A number of reads were removed as described by Johnson et al. [63]: (i) clonal reads, (ii) reads with low mapping quality score (<20), (iii) reads with more than 2 % mismatch to converted reference over the alignment length, (iv) reads mapping on the forward and reverse strand of the bisulfite converted genome, (v) read pairs not mapped at the expected distance based on library insert size, and (vi) read pairs that mapped in the wrong direction.

Additional files

Additional file 1: Table S1. List of the individuals and tissue samples in the WGBS datasets analyzed from the NIH Roadmap Epigenomics project.

Additional file 2: Table S2. List of the subset of probes from the 450 K that are consistent across tissues of the same individual but that vary between individuals, providing a basis for distinguishing individuals on the basis of methylation in any tissue.

Abbreviations

450 K: Illumina Infinium HumanMethylation 450 K BeadChip Array; CpG: Cytosine-phosphate-guanine; CpH: Cytosine-phosphate-(non-guanine nucleotide); DAC: Data Access Committee; EMC: McGill Epigenome Mapping Centre; gDNA: Genomic DNA; GEO: Gene Expression Omnibus; GINA: US Genetic Information Nondiscrimination Act, 2008; IHEC: International Human Epigenome Consortium; NCBI: National Centre for Biotechnology Information; NIH: National Institute of Health; PGP: Personal Genome Project; SNP: Single nucleotide polymorphism; TCGA: The Cancer Genome Atlas; WGBS: Whole-genome bisulfite sequencing.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

SD, YJ, JW, RS and TP devised the study. SD conducted the ethical-legal research and policy analysis. WC performed the bioinformatics analyses. MC, SBu, GB and LR performed the initial acquisition and bioinformatics data processing for the whole-genome bisulfite sequencing, 450 k array and genotype array data from the McGill Epigenomics Mapping Centre datasets. SD, WC and TP drafted the manuscript, with revisions and scientific guidance from YJ, OA, PL, MR, PF, SBe, MH, HS, RS and JW. All authors approved the final manuscript.

Acknowledgments

We thank Dr Mike Pazin for his valuable comments on the manuscript. We also thank members of the IHEC Bioethics Workgroup for helpful discussion of this work and Katie Saulnier for research assistance. SD, WC, YJ, MC, SBu, GB and TP receive grant support from the Canadian Institutes of Health Research (EP1-120608; EP2-120609). SD is also supported by the Canada Research Chair in Law and Medicine. WC is supported by a fellowship from the Fonds de Recherche du Quebec (FRSQ-30270). YJ receives grant support from the Fonds de Recherche du Quebec (FRSQ-24463). TP holds a Canada Research Chair. OA and RS receive grant support for genome-wide epigenomic studies from the European Union in the framework of the BLUEPRINT (HEALTH-F5-2011-282510) and SAME (EU, 57-1.3-10; Interreg) projects, the German Ministry of Science and Education (BMBF) in the framework of the ICGC MMML-Seq project (01KU1002A), the German Center for Lung Research (82DZL00105), the Imprinting-Network (01GM1114E; 01GM1513D) and the MMML-MYC-SYS project (036166B). LR acknowledges the Swedish Research Council and the Knut and Alice Wallenberg Foundation. PF acknowledges that the research leading to these results has received funding from the European Union's Seventh Framework Programme (FP7/2007-2013) under grant agreement 282510-BLUEPRINT and from the European Molecular Biology Laboratory. SBe was supported by EU-FP7 projects EpiGeneSys (257082) and BLUEPRINT (282510) and by a Royal Society Wolfson Research Merit Award (WM100023).

Author details

¹Centre of Genomics and Policy, Department of Human Genetics, McGill University, Montreal, QC H3A 0G1, Canada. ²Department of Human Genetics, McGill University and Genome Quebec Innovation Centre, Montreal, QC H3A 0G1, Canada. ³Institute of Human Genetics, University Hospital Schleswig-Holstein, Campus Kiel & Christian-Albrechts-University Kiel, 24105 Kiel, Germany. ⁴Saarland University, 66123 Saarbrücken, Germany. ⁵Institute for Bioethics, Health Policy and Law, University of Louisville School of Medicine, Louisville, KY 40202, USA. ⁶Department of Medical Sciences, Science for Life Laboratory, Uppsala University, SE-751 85 Uppsala, Sweden. ⁷European Molecular Biology Laboratory, European Bioinformatics Institute,

Wellcome Genome Campus, Hinxton, Cambridge CB10 1SD, UK. ⁸Medical Genomics, UCL Cancer Institute, University College London, London WC1E 6BT, UK. ⁹Centre for High-Throughput Biology, University of British Columbia and Canada's Michael Smith Genome Sciences Centre, BC Cancer Agency, Vancouver, British Columbia V5Z 4S6, Canada. ¹⁰Department of Molecular Biology, RIMLS, Faculty of Science, Radboud University, 6500 HB Nijmegen, The Netherlands.

Published online: 17 July 2015

References

- Grundberg E, Meduri E, Sandling JK, Hedman AK, Keildson S, Buil A, et al. Global analysis of DNA methylation variation in adipose tissue from twins reveals links to disease-associated variants in distal regulatory elements. *Am J Hum Genet.* 2013;93:876–90.
- Erlach Y, Narayanan A. Routes for breaching and protecting genetic privacy. *Nat Rev Genet.* 2014;15:409–21.
- Lin Z, Owen AB, Altman RB. Genetics. Genomic research and human subject privacy. *Science.* 2004;305:183.
- Lee HY, Park MJ, Yoo JE, Chung U, Han GR, Shin KJ. Selection of twenty-four highly informative SNP markers for human identification and paternity analysis in Koreans. *Forensic Sci Int.* 2005;148:107–12.
- Joly Y, Zeps N, Knoppers BM. Genomic databases access agreements: legal validity and possible sanctions. *Hum Genet.* 2011;130:441–9.
- Greenbaum D, Sboner A, Mu XJ, Gerstein M. Genomics and privacy: implications of the new reality of closed data for the field. *PLoS Comput Biol.* 2011;7:e1002278.
- National Institutes of Health. NIH Genomic Data Sharing Policy (NOT-OD-14-124). 2014. <http://grants.nih.gov/grants/guide/notice-files/NOT-OD-14-124.html>. Accessed 1 July 2015.
- Ohno-Machado L, Bafna V, Boxwala AA, Chapman BE, Chapman WW, Chaudhuri K, et al. iDASH: integrating data for analysis, anonymization, and sharing. *J Am Med Inform Assoc.* 2012;19:196–201.
- Nietfeld JJ, Sugarman J, Litton JE. The Bio-PIN: a concept to improve biobanking. *Nat Rev Cancer.* 2011;11:303–8.
- Wolfson M, Wallace SE, Masca N, Rowe G, Sheehan NA, Ferretti V, et al. DataSHIELD: resolving a conflict in contemporary bioscience - performing a pooled analysis of individual-level data without sharing the data. *Int J Epidemiol.* 2010;39:1372–82.
- Data Levels and Data Types. The Cancer Genome Atlas Data Portal. <https://tcga-data.nci.nih.gov/tcga/tcgaDataType.jsp>. Accessed 1 July 2015.
- Kilpinen H, Waszak SM, Gschwind AR, Raghav SK, Witwicki RM, Orioli A, et al. Coordinated effects of sequence variation on DNA binding, chromatin structure, and transcription. *Science.* 2013;342:744–7.
- Degner JF, Pai AA, Pique-Regi R, Veyrieras JB, Gaffney DJ, Pickrell JK, et al. DNase I sensitivity QTLs are a major determinant of human expression variation. *Nature.* 2012;482:390–4.
- McVicker G, van de Geijn B, Degner JF, Cain CE, Banovich NE, Raj A, et al. Identification of genetic variants that affect histone modifications in human cells. *Science.* 2013;342:747–9.
- Lin X, Sun D, Rodriguez B, Zhao Q, Sun H, Zhang Y, et al. BSeQC: quality control of bisulfite sequencing experiments. *Bioinformatics.* 2013;29:3227–9.
- Hansen KD, Langmead B, Irizarry RA. BSmooth: from whole genome bisulfite sequencing reads to differentially methylated regions. *Genome Biol.* 2012;13:R83.
- Jin SG, Kadam S, Pfeifer GP. Examination of the specificity of DNA methylation profiling techniques towards 5-methylcytosine and 5-hydroxymethylcytosine. *Nucleic Acids Res.* 2010;38:e125.
- Guo JU, Su Y, Zhong C, Ming GL, Song H. Hydroxylation of 5-methylcytosine by TET1 promotes active DNA demethylation in the adult brain. *Cell.* 2011;145:423–34.
- Liu Y, Siegmund KD, Laird PW, Berman BP. Bis-SNP: combined DNA methylation and SNP calling for Bisulfite-seq data. *Genome Biol.* 2012;13:R61.
- Barturen G, Rueda A, Oliver JL, Hackenberg M. MethylExtract: high-quality methylation maps and SNV calling from whole genome bisulfite sequencing data. *F1000Res.* 2013;2:217.
- Bernstein BE, Stamatoyannopoulos JA, Costello JF, Ren B, Milosavljevic A, Meissner A, et al. The NIH roadmap epigenomics mapping consortium. *Nat Biotechnol.* 2010;28:1045–8.

22. McGill Epigenomics Mapping Centre Portal. <http://epigenomesportal.ca>. Accessed 1 July 2015.
23. Allum F, Shao X, Guénard F, Simon MM, Busche S, Caron M, et al. Characterization of functional methylomes by next-generation capture sequencing identifies novel disease associated variants. *Nat Commun*. 2015;6:7211.
24. Busche S, Ge B, Vidal R, Spinella JF, Sailour V, Richer C, et al. Integration of high-resolution methylome and transcriptome analyses to dissect epigenomic changes in childhood acute lymphoblastic leukemia. *Cancer Res*. 2013;73:4323–36.
25. Grundberg E, Small KS, Hedman AK, Nica AC, Buil A, Keildson S, et al. Mapping cis- and trans-regulatory effects across multiple tissues in twins. *Nat Genet*. 2012;44:1084–9.
26. Malin B, Sweeney L. Re-identification of DNA through an automated linkage process. *Proc AMIA Symp*. 2001:423–7.
27. Sobrido MJ, Coppola G, Oliveira J, Hopfer S, Geschwind DH. Primary familial brain calcification. In: Pagon RA, Adam MP, Ardinger HH, editors. *GeneReviews*® [Internet]. Seattle (WA): University of Washington; 1993–2015.
28. Consumer Partnership for eHealth. Protecting sensitive health information in the context of health information technology. June 2010. <http://www.nationalpartnership.org/research-library/health-care/HIT/protecting-sensitive-health.pdf>. Accessed 1 July 2015.
29. Kaback MM, Desnick RJ. Hexosaminidase A Deficiency. In: Pagon RA, Adam MP, Ardinger HH, editors. *GeneReviews*® [Internet]. Seattle (WA): University of Washington; 1993–2015.
30. Laberge AM, Michaud J, Richter A, Lemyre E, Lambert M, Brais B, et al. Population history and its impact on medical genetics in Quebec. *Clin Genet*. 2005;68:287–301.
31. Weidner CI, Lin Q, Koch CM, Eisele L, Beier F, Ziegler P, et al. Aging of blood can be tracked by DNA methylation changes at just three CpG sites. *Genome Biol*. 2014;15:R24.
32. Kasowski M, Kyriazopoulou-Panagiotopoulou S, Grubert F, Zaugg JB, Kundaje A, Liu Y, et al. Extensive variation in chromatin states across humans. *Science*. 2013;342:750–2.
33. Benitez K, Malin B. Evaluating re-identification risks with respect to the HIPAA privacy rule. *J Am Med Inform Assoc*. 2010;17:169–77.
34. Sankar P, Cho MK. Genetics. Toward a new vocabulary of human genetic variation. *Science*. 2002;298:1337–8.
35. Caulfield T, Fullerton SM, Ali-Khan SE, Arbour L, Burchard EG, Cooper RS, et al. Race and ancestry in biomedical research: exploring the challenges. *Genome Med*. 2009;1:8.
36. National Institutes of Health. NIH policy and guidelines on the inclusion of women and minorities as subjects in clinical research; 2001. <http://grants.nih.gov/grants/guide/notice-files/NOT-OD-02-001.html>. Accessed 1 July 2015.
37. International HapMap Consortium. Integrating ethics and science in the International HapMap Project. *Nat Rev Genet*. 2004;5:467–75.
38. Smart A, Tutton R, Martin P, Ellison GT, Ashcroft R. The standardization of race and ethnicity in biomedical science editorials and UK biobanks. *Soc Stud Sci*. 2008;38:407–23.
39. Gitschier J. Inferential genotyping of Y chromosomes in Latter-Day Saints founders and comparison to Utah samples in the HapMap project. *Am J Hum Genet*. 2009;84:251–8.
40. Schadt EE, Woo S, Hao K. Bayesian method to predict individual SNP genotypes from gene expression data. *Nat Genet*. 2012;44:603–8.
41. Miura F, Enomoto Y, Dairiki R, Ito T. Amplification-free whole-genome bisulfite sequencing by post-bisulfite adaptor tagging. *Nucleic Acids Res*. 2012;40:e136.
42. Byun HM, Siegmund KD, Pan F, Weisenberger DJ, Kanel G, Laird PW, et al. Epigenetic profiling of somatic tissues from human autopsy specimens identifies tissue- and individual-specific DNA methylation patterns. *Hum Mol Genet*. 2009;18:4808–17.
43. Chen YA, Lemire M, Choufani S, Butcher DT, Grafodatskaya D, Zanke BW, et al. Discovery of cross-reactive probes and polymorphic CpGs in the Illumina Infinium HumanMethylation450 microarray. *Epigenetics*. 2013;8:203–9.
44. US-EU Safe Harbor Framework Documents. <http://www.export.gov/safeharbor/eu/>. Accessed 3 July 2015.
45. Milius D, Dove ES, Chalmers D, Dyke SO, Kato K, Nicolas P, et al. The International Cancer Genome Consortium's evolving data-protection policies. *Nat Biotechnol*. 2014;32:519–23.
46. Joly Y, Ngueng Feze I, Simard J. Genetic discrimination and life insurance: a systematic review of the evidence. *BMC Med*. 2013;11:25.
47. Caulfield T, Evans J, McGuire A, McCabe C, Bubela T, Cook-Deegan R, et al. Reflections on the cost of 'low-cost' whole genome sequencing: framing the health policy debate. *PLoS Biol*. 2013;11:e1001699.
48. Godard B, Pratte A, Dumont M, Simard-Lebrun A, Simard J. Factors associated with an individual's decision to withdraw from genetic testing for breast and ovarian cancer susceptibility: implications for counseling. *Genet Test*. 2007;11:45–54.
49. Phoenix Strategic Perspectives Inc. Survey of Canadians on privacy-related issues. Office of the Privacy Commissioner of Canada Ottawa, Canada. 2013. https://www.priv.gc.ca/information/por-rop/2013/por_2013_01_e.asp. Accessed 1 July 2015.
50. Keogh LA, van Vliet CM, Studdert DM, Maskiell JA, Macrae FA, St John DJ, et al. Is uptake of genetic testing for colorectal cancer influenced by knowledge of insurance implications? *Med J Aust*. 2009;191:255–8.
51. Geelen E, Horstman K, Marcelis CL, Doevendans PA, Van Hoyweghen I. Unravelling fears of genetic discrimination: an exploratory study of Dutch HCM families in an era of genetic non-discrimination acts. *Eur J Hum Genet*. 2012;20:1018–23.
52. Church GM. The personal genome project. *Mol Syst Biol*. 2005;1:2005.0030.
53. Lunshof JE, Chadwick R, Vorhaus DB, Church GM. Science and society - from genetic privacy to open consent. *Nat Rev Genet*. 2008;9:406–11.
54. Ball MP, Bobe JR, Chou MF, Clegg T, Estep PW, Lunshof JE, et al. Harvard Personal Genome Project: lessons from participatory public research. *Genome Med*. 2014;6:10.
55. Joly Y, Burton H, Knoppers BM, Feze IN, Dent T, Pashayan N, et al. Life insurance: genomic stratification and risk classification. *Eur J Hum Genet*. 2014;22:575–9.
56. Rothstein MA, Cai Y, Marchant GE. The ghost in our genes: legal and ethical implications of epigenetics. *Health Matrix Clevel*. 2009;19:1–62.
57. Rothstein MA. Epigenetic exceptionalism. *J Law Med Ethics*. 2013;41:733–6.
58. U.S. Equal Employment Opportunity Commission. The Genetic Information Nondiscrimination Act (GINA) of 2008. Section 201(7). <http://www.eeoc.gov/laws/statutes/gina.cfm>. Accessed 1 July 2015.
59. Joly Y, Dove ES, Knoppers BM, Bobrow M, Chalmers D. Data sharing in the post-genomic world: the experience of the International Cancer Genome Consortium (ICGC) Data Access Compliance Office (DACO). *PLoS Comput Biol*. 2012;8:e1002549.
60. Philibert RA, Terry N, Erwin C, Philibert WJ, Beach SRH, Brody GH. Methylation array data can simultaneously identify individuals and convey protected health information: an unrecognized ethical concern. *Clin Epigenet*. 2014;6:28.
61. NIH Roadmap Epigenomics Project GEO Repository. ftp://ftp.ncbi.nlm.nih.gov/pub/geo/DATA/roadmapepigenomics/by_experiment/Bisulfite-Seq/. Accessed 1 July 2015.
62. Lister R, Pelizzola M, Dowen RH, Hawkins RD, Hon G, Tonti-Filippini J, et al. Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature*. 2009;462:315–22.
63. Johnson MD, Mueller M, Game L, Aitman TJ. Single nucleotide analysis of cytosine methylation by whole-genome shotgun bisulfite sequencing. *Curr Protoc Mol Biol*. 2012;Chapter 21:Unit21.23.